paper_draft

DM

Title

Statistical Clinical Phenotype and Patient Survival Analysis in Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies (NEDMABA)

Abstract

A recently described, rare genetic condition known as Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies (NEDMABA) has been identified in children with in children with bi-allelic loss-of-function variants in *SMPD4*.

Few case reports have been described to date, with a diverse and severe clinical phenotype in most reported cases, the progression of this condition is not well understood.

A gap exists in the understanding the associations of the heterogenous features present in the clinical phenotype and the expected survival probabilities of indivudals with this condition. This is driven in part to the paucity of analysis-ready data on subjects.

This research aims to collate and standardise available case reports, to analyse and identify meaningful clusters in the clinical phenotype and the quantify the survival probability function for children with NEDMABA.

To overcome the challenge of sparse, multidimensional data on very few subjects, we employ Multiple Correspondece Analysis as a dimension reduction technique, which is then subject to heirarchical clustering and interpretation. To quantify more accurate survival probabilities, kaplan-meier estimation is formulated to account for censoring in the data.

Several distinct clusters of clinical features were discovered in the current case reports which can help explain the variation present in the clinical phenotype for children with this condition. Furthermore, survival probability for those affected declines sharply in early infancy, but exhibits a wide range of outcomes provisionally associated with variant type.

We are able to identify the classic phenotype for this condition, as well as two other auxillary feature clusters. The first relates findings of vocal cord paralysis and swallowing dysfunction, the other relates to more complex brain anomalies. In addition, several outlying or non-typical features are able to be classified. Despite strong indications of differences in survival outcomes based on variant type, interpretation of these results are guarded based on very low sample sizes.

Ultra rare conditions are a challenge for clinicians and families, with often limited information on diverse outcomes. Despite challenges with sparse and inconsistent data, analytical techniques to better describe associations and trajectories of indivudals with these conditions can provide clinicians and genetic counsellors with enhanced information to aide in decision making and support for families.

Introduction

Neurodevelopment disorders are a diverse and heterogeneous group of conditions impacting the development of the nervous system, brain function, physical development, emotional development and learning ability. A recently identified condition known as Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies (NEDMABA) (MIM:618622) has been described in children with bi-allelic loss-of-function variants in *SMPD4* (Magini et al. 2019). Sphingomyelinases such as *SMPD4* play an important celluar role by hydrolyzing sphingomyelin into ceramide and phosphorylcholine. *SMPD4* specifically encodes one of 4 neutral sphinogmyelinases, nSMase3 (MIM: 610457).

So far, only approximately 30 cases of this rare disorder have been reported across over 12 families. Clinical phenotype data of these cases is largely heterogeneous with severe neurological complications and early demise a key feature. This makes the identification, diagnosis and ongoing management of cases a challenge for patients, their families and medical practitioners.

Ravenscroft et al. (2021) performed a study over 190 probands with a diagnosis of arthrogryposis multiplex congenita, distal arthrogryposis, fetal akinesia deformation sequence or multiple pterygium syndrome. This study identified a novel missense variation in SMPD4 which at the time of whole exome sequencing was not well described in the literature. The impacted family from Melbourne exhibited features involving arthrogryposis multiplex congenita, complex brain malformations, small for gestation age and hypoplasia of the corpus callosum. In two of the three related cases were additional features of microcephaly, congenital encephalopathy, cerebellar malformation and hypoplasia and hypomyelination.

Monies et al. (2019) described two families with different homozygous truncating variants in SMPD4 as a syndrome of skeletal dysplasia with cerebella atrophy. The phenotypes described involved bilateral clenched hands, talipes, IUGR, partial absence of corpus callosum with family history of three neonatal deaths with similar features.

A detailed study by Magini et al. (2019) involved 12 unrelated familes with 32 individuals (21 with detailed clinical information). Hallmark presentations were of microcephaly, simplified gyration, hypomyelination, thin corpus callosum, mild cerebellar hypoplasia, brainstem hypoplasia, congenital arthrogryposis, diabetes mellitus, heart disease, severe encephalopathy and respiratory problems often leading to early demise. Despite this being the largest cohort studied, the clinical features and survival times among participants varied greatly. Three missense changes were noted in the study with affected children in these families often showing a milder presentation suggestive of possible residual function. In these cases individuals were able to develop independent motor skills, have mild intellectual disability and arthrogryposis without evidence of simplified gyral patterns on brain MRI. Other patients with truncating variants are shown to have more severe presentations, while a range of additional significant clinical features were reported involving dysmorphic facial features, seizure, vocal cord paralysis and hearing impairment.

A recent case study from China is described in Ji et al. (2022) involving a girl presenting in infancy with intrauterine growth restriction, microcephaly, postnatal developmental delay, arthrogryposis, hypertonicity, seizure, and hypomyelination on brain magnetic resonance imaging. Ji et al. (2022) argues the actual prevalence of this condition, based on the gene carrier rate is understated and may be attributable to typical symptoms of NEDMABA being non-specific, providing a diagnostic challenge in a clinical setting.

A recent study by Bijarnia-Mahay et al. (2022) presents the case of a 22-month old girl presenting with the typical phenotype of neurodevelopmental delay, prenatal onset growth failure, arthrogryposis, microcephaly and brain anomalies including severe hypomyelination, simplified gyral pattern and hypoplasia of corpus callosum and brainstem. Notably, there is also additional non-typical clinical findings of nystagmus and visual impairment secondary to macular dystrophy and retinal pigment epithelial stippling at posterior pole.

Ji et al. (2022) reports parallels with two cases showing the same homozygous null variant. An individual reported in Monies et al. (2019) presented with distinct symptoms of brain atrophy and skeletal dysplasia whereas the case in Magini et al. (2019) with the same variant exhibited more typical clinical features.

Further work to collate and analyse data relating to NEDMABA is challenging. While Magini et al. (2019) cataloged a detailed clinical phenotype data set as a supplementary data set to their study, these data are not suitable for statistical analysis as presented. Many of the clinical features are represented as free text descriptions and a variety of non-standard terminologies are applied between cases, making machine interpretation of the data difficult. Other studies (Ji et al. 2022; Bijarnia-Mahay et al. 2022) present only written case reports with some tabulated summaries. While larger studies (Ravenscroft et al. 2021; Monies et al. 2019) focus more on genetic data, with less detail included on the clinical presentation of individuals in the cohort.

To overcome these challenges, Wickham (2014) proposes a data structure know as ‘tidy’ data which organises each observation in a row, with each feature as its own column and each value in just one cell. In this case detailed text-based data can be tranformed to high dimensional binary indicators of key clinical features. This structure is conducive to effective statistical analysis and integrates deliberately with the *tidyverse* (Wickham et al. 2019) collection of packages for data analysis in R (R Core Team 2022).

Further analysis to better describe rare genetic conditions was explored in Díaz-Santiago et al. (2020) in the context of large scale genotype-phenotype analysis. The authors argue patients with rare disorders (often in small samples) often present with varied symptoms that do not match exactly with the described phenotype. When considering this situation of low sample size data with many features, methods in the field of multivariate statistical analysis are commonly deployed. Here the aim is to find substructure in the data, or a simple representation of the multi-dimensional space. Methods, such as Multiple Correspondence Analysis (Le Roux and Rouanet 2010) are appropriate for high dimensional data comprised of binary indicators. This technique is cited in many studies in the application of dimension reduction and clustering of comorbidities and phenotypes from disease (Han, Benseler, and Tyrrell 2018; Costa et al. 2013).

When the focus is on survival or mortality, methods in statistical survival analysis are well suited and commonly used in the field of genomics (Chen, Sun, and Hoshida 2014). These methods in particular provide a mechanism for dealing with right-censoring, a common feature of clinical studies where the clinical outcome of interest may not be known by the end of the study period. Within the context of examining a single condition, Crowe et al. (2020) investigated comorbidity phenotypes and mortality risk in ischaemic heart disease patients. Here latent class analysis was performed to identify a small number of clusters in the patients which could then be included in survival analysis techniques to better understand the substructre and outcome of patients of this condition.

An open research question exists around the clinical pathway and survival time of children exhibiting bi-allelic loss-of-function variants in SMPD4. With current research highlighting a diverse and heterogeneous phenotype, the relationship and correlations between these diverse features is not well understood. Furthermore, the survival time of affected children is not well understood despite early-demise featuring as a severe outcome. While a connection has been highlighted between some missense variants and a milder presentation with longer survival, this hypothesis has not been analysed further.

This research has three key aims. First, to collate and transform the variety of early case reports and studies on clinical phenotype data for this novel variant into an analysis-ready *tidy* data set. Secondly, to conduct analysis on the associations between both patients and their clinical features. Finally, to statistically quantify the expected survival time of children with NEDMABA based on the current case reports.

This will be the first in-depth analysis of early studies of this novel variant, which aims to produce statistical findings to better describe and understand this condition. This research is significant as it will assist clinicians understand the typical and non-typical presentations of this challenging new condition. In addition, genetic counselling of families with affected children will benefit from enhanced analysis on outcomes of existing reported cases.

Methods

SMPD4 Data Package

The data contained in Magini et al. (2019) "Summary of SMPD4-Related Clinical Phenotype" is an excel spreadsheet tabulating each of the 21 individuals from the study with clinical details recorded. The file has 21 columns (one for each individual), and 64 rows (one for each phenotype or clinical remark).

The data were read into R (R Core Team 2022) without any changes, so as to preserve the reproducibility of the data transformation steps.

The data were transposed to ensure it could be presented as *tidy* formatted data (Wickham 2014). This requires:

1. Each variable (clinical phenotype) forms a column.
2. Each observation (individual) forms a row.
3. Each type of observational unit forms a table (every cell has just one item)

In many cases, key clinical information was entered as free-text descriptions, which rendered any attempt of meaningful analysis impractical (Table 1). In these cases, the text was tokenised by separating the list of clinical observations at each comma and forming a binary indicator column noting its presence '1' or absence '0' (Table 2).

Table 1: Example of non-tidy free-text descriptions of features

	Family 1- Individual 1	Family 1- Individual 4
Facial dysmorphisms	short palpebral fissures, large ears, simple helices, smooth philtrum, thin lips, bilateral simian creases	short palpebral fissures, receding forehead, thin upper lip

Table 2: Example of tidy formatted data where individuals are transposed into rows and text into binary indicators

id	short palpebral fissures	large ears	simple helices	receding forehead	…
Family 1-Individual 1	1	1	0	…	
Family 1-Individual 4	1	0	0	1	…

In the case where two variables were formed from phenotypes that are considered to be synonymous, these were merged into one indicator column to prevent duplication e.g. {bilateral_cleft_lips, bilateral_cleft_lip, cleft_lip_b_l}.

Some data type conversion and categorical level standardisation was performed to ensure the data were in consistent and appropriate data types. For example, 'Gender' was not consistently coded, and 'Birth Weight' was encoded as a text string rather than a more useful numeric format. (Table 3).

Table 3: Example of inconsistent coding or sub-optimal data types

Gender	Birth Weight
male	2175 grams (- 2.5 SD)
female	2045 g (-3 SD)
Female	2300 gram (-2 SD)
Female fetus	n.a.

The final dataset consisted of 21 observations (one per individual) and 152 variables (one per clinical feature).

This format was preserved and other case studies identified in the literature (Ravenscroft et al. 2021; Monies et al. 2019; Bijarnia-Mahay et al. 2022; Ji et al. 2022) were manually entered to conform to this template to allow the data to be combined for further analysis.

These data sets were packaged into an R Package called SMPD4 (Marchiori 2022) in order to allow for reproducibility and sharing. This can be downloaded and installed from github at https://github.com/deanmarchiori/SMPD4.

Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) is an analogy to Principle Component Analysis (PCA) for categorical data (Le Roux and Rouanet 2010).

We let a data set $\mathbf{X}$ be comprised of a set of individuals I and a set of features Q such that the q_{th} feature has K_q levels. The sum of all categories $K = \sum_{q=1}^{Q} K_q$ defines the dimensionality of $\mathbf{X}$ as an $I \times K$ matrix.

Taking $\delta_{ik} = 1$ if subject i has feature k and $\delta_{ik} = 0$ if the subject does not, we are left with the completely disjnuctive table $\mathbf{X} = I \times K$ of $\{0, 1\}$. Letting the sum of all entries of $\mathbf{X}$ be N, we have $\mathbf{Z} = N^{-1}\mathbf{X}$.

We can introduce two diagonal matricies $\mathbf{D_r} = diag(\mathbf{r})$ and $\mathbf{D_c} = diag(\mathbf{c})$ where $\mathbf{r}$ and $\mathbf{c}$ are the vectors of row sums and column sums of $\mathbf{Z}$ respectively.

Computing MCA involves taking the Singular Value Decomposition of:

$$\mathbf{M} = \mathbf{D_r}^{-\frac{1}{2}}(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathbf{T}})\mathbf{D_c}^{-\frac{1}{2}} = \mathbf{P}\Delta\mathbf{Q}^{\mathbf{T}}$$

where Δ are the singular values and $\Lambda = \Delta^2$ is the matrix of eigenvalues.

This results in the row and column factor scores respectively as:

$$\mathbf{F} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{P}\Delta$$

$$\mathbf{G} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{Q}\Delta$$

The combined clinical phenotype data was subsetted to include only those features that were encoded as a binary indicator variable.

Only those features that appeared in more than one case were included to minimise the influence of non-related features in this analysis.

This resulted in a dataset $X_{28 \times 61}$ indicator matrix of 61 clinical features across 28 individuals.

[NEED TO INCLUDE CONTRIBUTION]

Multiple Correspondence Analysis was then computed using the *FactoMineR* R package (Lê, Josse, and Husson 2008).

Cluster Analysis

[insert clvalid details]

Survival Analysis

To model survival probability for all subjects in the combined data, the data are subsetted to include *survival_time* which is either the number of days the individual survived for, or the age in days at last follow up, and *deceased* a numeric indicator which is equal to 1 if the subject is deceased and 0 otherwise.

The Kaplan-Meier estimator (Kaplan and Meier 1958) is used to estimate the survival function of the data.

The estimator is calcualted as:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

where t_i is some event time, d_i represents the number of events (here deceased subjects) and n_i indicating the indivudals known to have survived up to t_i

The baseline estimator of all individuals was calculated using the *survival* package in R (Terry M. Therneau and Patricia M. Grambsch 2000).

Next the Kaplan-Meier estimator was calculated, [ON WHAT?] stratified by the reported type of Single nucleotide polymorphisms (SNP) reported in the literature. A test was performed to detect differences in the survival curves using methods from Harrington and Fleming (1982), again implemented in the *survival* package.

Finally, a Cox proportional hazards regression model (Andersen and Gill 1982) was fit on survival time with the variant type as the sole covariate. This was to detect and measure the Hazard ratio of variant type on the survival time in order to test the hypothesis that misense variant types exhibit significantly greater survival compared to other variant types.

Results

MCA

The initial results of the MCA were inspected with a bi-plot and identified two individuals (17, 18) from the data that were clear outliers and sat significatly beyound the 95% confidence ellipse. These individuals represented two twins from Magini et al. (2019) who were described with a much milder phenotype and wer subject to a missense variation in SMPD4. Further analysis of this influence is conducted below. For now, these subjects were removed to allow for more meaningful analysis.

Repeating MCA for the remaining subjects shows the first dimension ($\lambda_! = 0.18$) accounts for 17.88% of the variance in the data with the first 4 dimensions of the MCA analysis accounting for 54% of the variance.

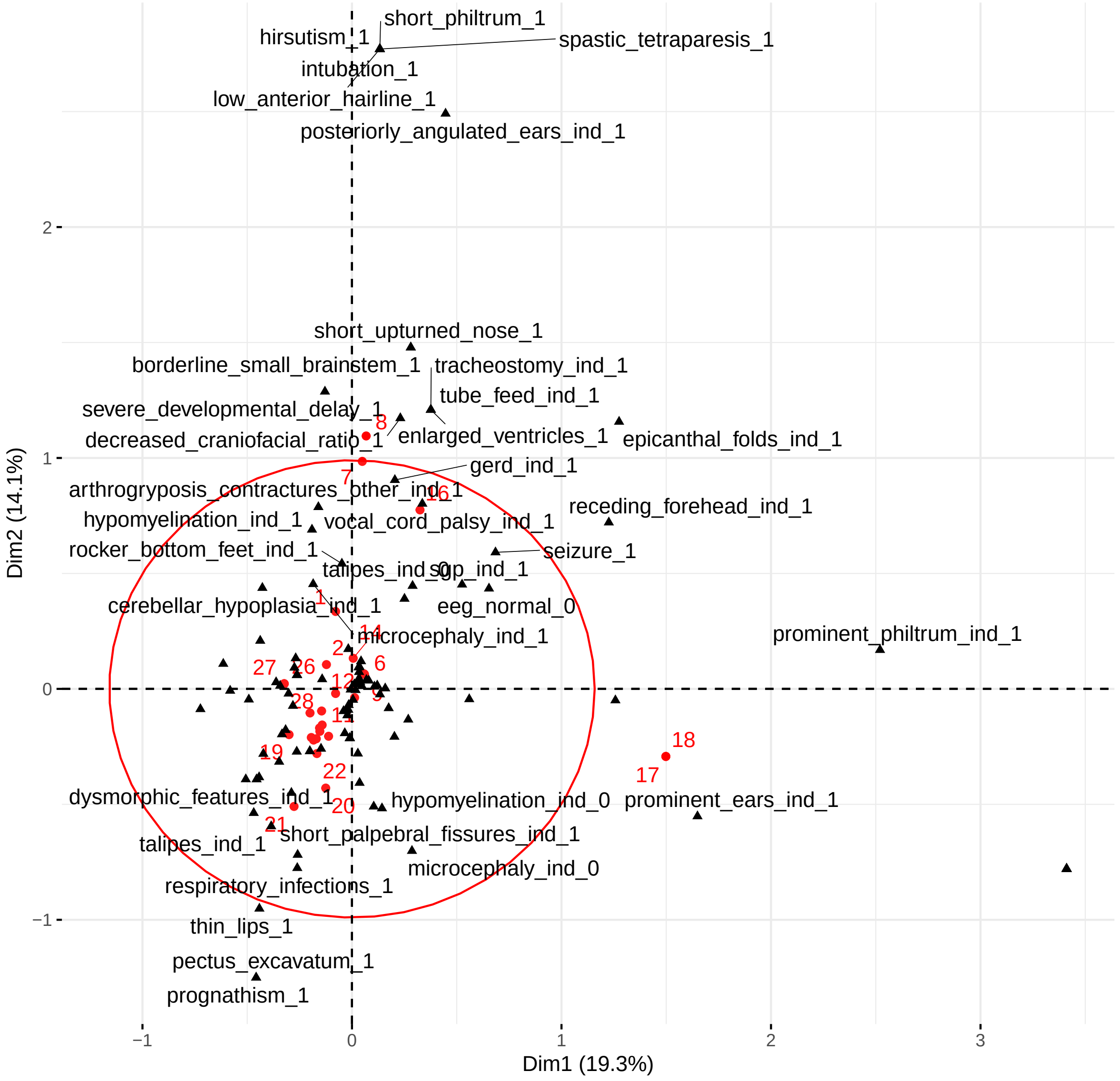


Figure 1: A bi-plot of the first two principal axes with grey cicles indicating the individual's projected position and black triangles indicaating the variables

Table 4: eigenvalues and percentage of variance explained by the MCA principal axes

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.18	17.88	17.88
dim 2	0.14	13.81	31.70
dim 3	0.12	12.32	44.01
dim 4	0.10	9.66	53.67
dim 5	0.07	7.49	61.16
dim 6	0.06	5.97	67.13
dim 7	0.05	4.90	72.03
dim 8	0.04	4.20	76.23
dim 9	0.04	4.08	80.31
dim 10	0.03	3.47	83.78
dim 11	0.03	2.92	86.70
dim 12	0.02	2.49	89.20
dim 13	0.02	2.15	91.34
dim 14	0.02	1.82	93.16
dim 15	0.02	1.65	94.81
dim 16	0.01	1.12	95.93
dim 17	0.01	1.00	96.93
dim 18	0.01	0.93	97.85
dim 19	0.01	0.67	98.53
dim 20	0.01	0.63	99.16
dim 21	0.00	0.41	99.57
dim 22	0.00	0.31	99.88
dim 23	0.00	0.12	100.00
dim 24	0.00	0.00	100.00
dim 25	0.00	0.00	100.00

The *contribution* of the categories were ranked within each of the first four MCA dimensions See Figure 2.

The first dimension is dominated by highly specific facial dysmophisms such as epicanthal folds, short philtrum and low anteriour hairline. Also significant are seizure and developmental delay.

The second MCA dimension is categorised by feeding and respiratory dysfunction with vocal cord palsy, tracheostomy, tube feeding and GERD.

Examining the bi-plot of the first two principal axes Figure 3 shows contrasting sources of variation between conditions relating to airway and swallowing function and predominantly structural brain anomalies such as as cerebellar hypoplasia, agenesis of corpus collosum and development delay.

The third dimension is characterised by complex non-typical features such as patent ductus arteriosus, respiratory infections, small for gestational age and further facial features. These are contrasted by the fourth principal dimension which is typical for respiratory distress including stridor and hypercapnia.

Cluster Analysis

Internal cluster validation measures resulted in between 2 and 6 clusters as the optimal choice.

Proceeding with the optimal cluster selection, per the Silhouette coefficient of 6 clusters.

A scatterplot of the cluster results Figure 7 projected on just the first two dimensions highlights a core cluster of classic features of NEDMABA. A separate cluster is formed of airway and feeding related conditions involing vocal cord paralysis, tracheostomy, tube feeding and enlarged ventricles.

A cluster relating to structual brain anomalies is articulated with conditions such as cerebellar hypoplasia and agenesis of corpus collosum, along with small for gestational age.

Some smaller satellite clusters exists where highly distinctive features were associated with an indivual or family. Also a singleton clusters exist for features that are distinct or unusual such as posteriorly angulated ears.

Table 5: Optimal cluster numbers determined by four similar metrics measuring cluster analysis performance.

measure	metric	optimal cluster no.
Connectivity	7.50	2
Dunn	0.18	3
Dunn	0.18	4
Silhouette	0.45	6

Survival Analysis

The Kaplan-Meier estimator shows a range of survival times (0, 4562 days) with survival probability at time 0 of 0.82 (0.69, 0.98). Survival probability after one year at 0.403 (0.252, 0.644) with median survival of 150 days and survival probability at close to ten years of 0.24 (0.11, 0.52).

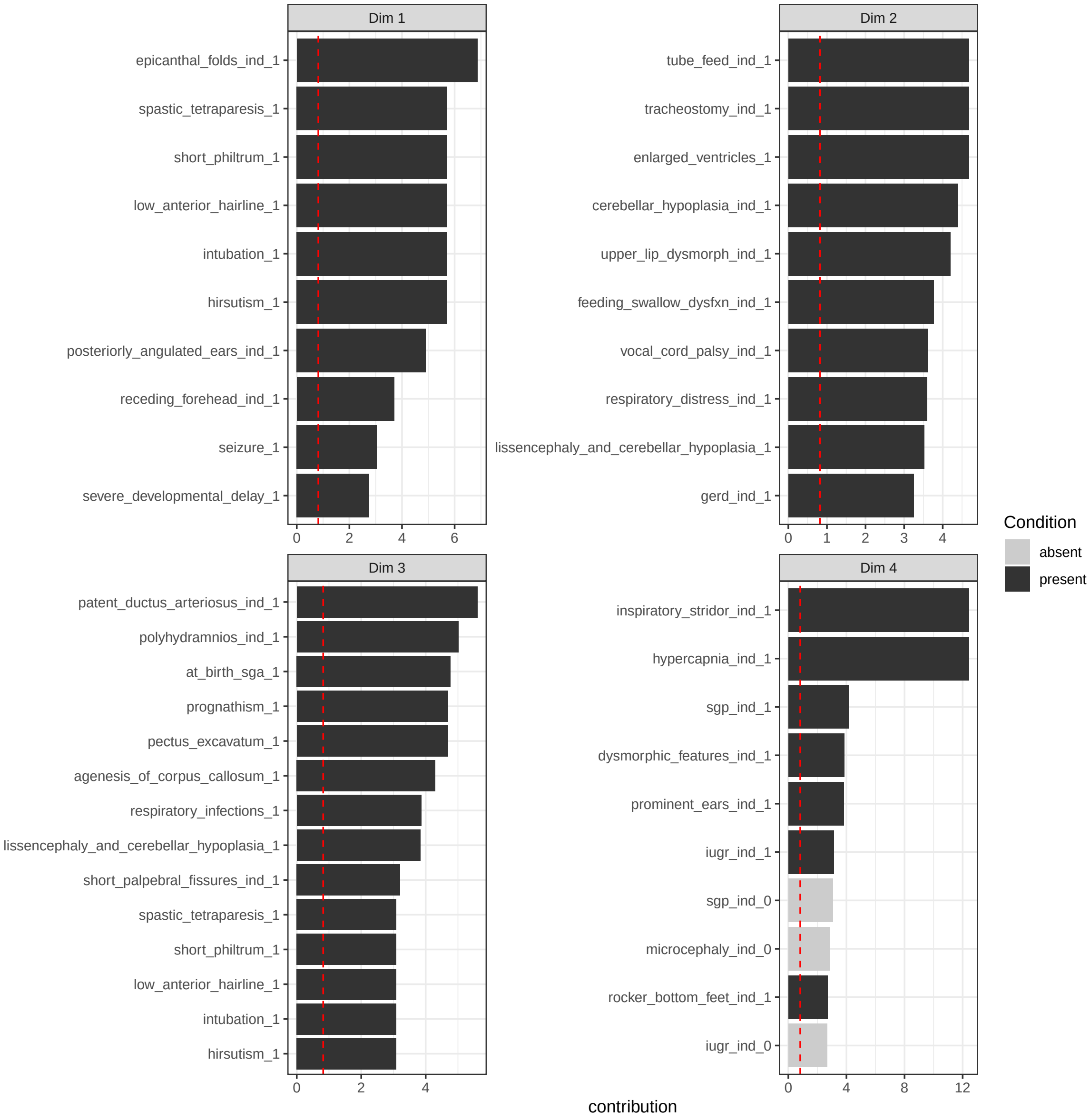


Figure 2: Top 20 feature categories from MCA analysis for the first four MCA dimensions. The red dashed line indicates the mean contribution value. Each feature may be present or absent as indicated by 1 or 0 respectively

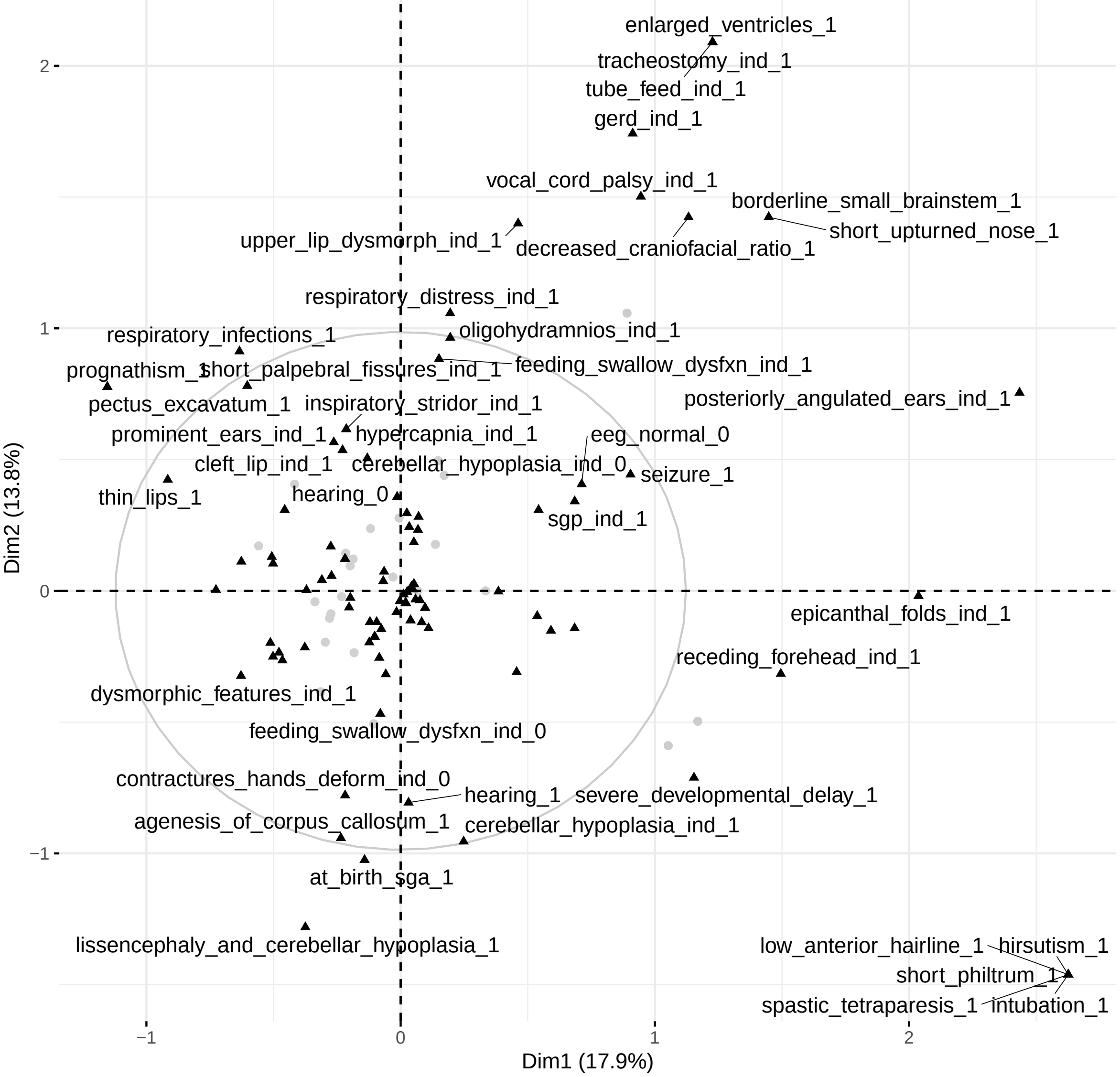


Figure 3: A bi-plot of the first two principal axes with grey cicles indicating the individual's projected position and black triangles indicaating the variables

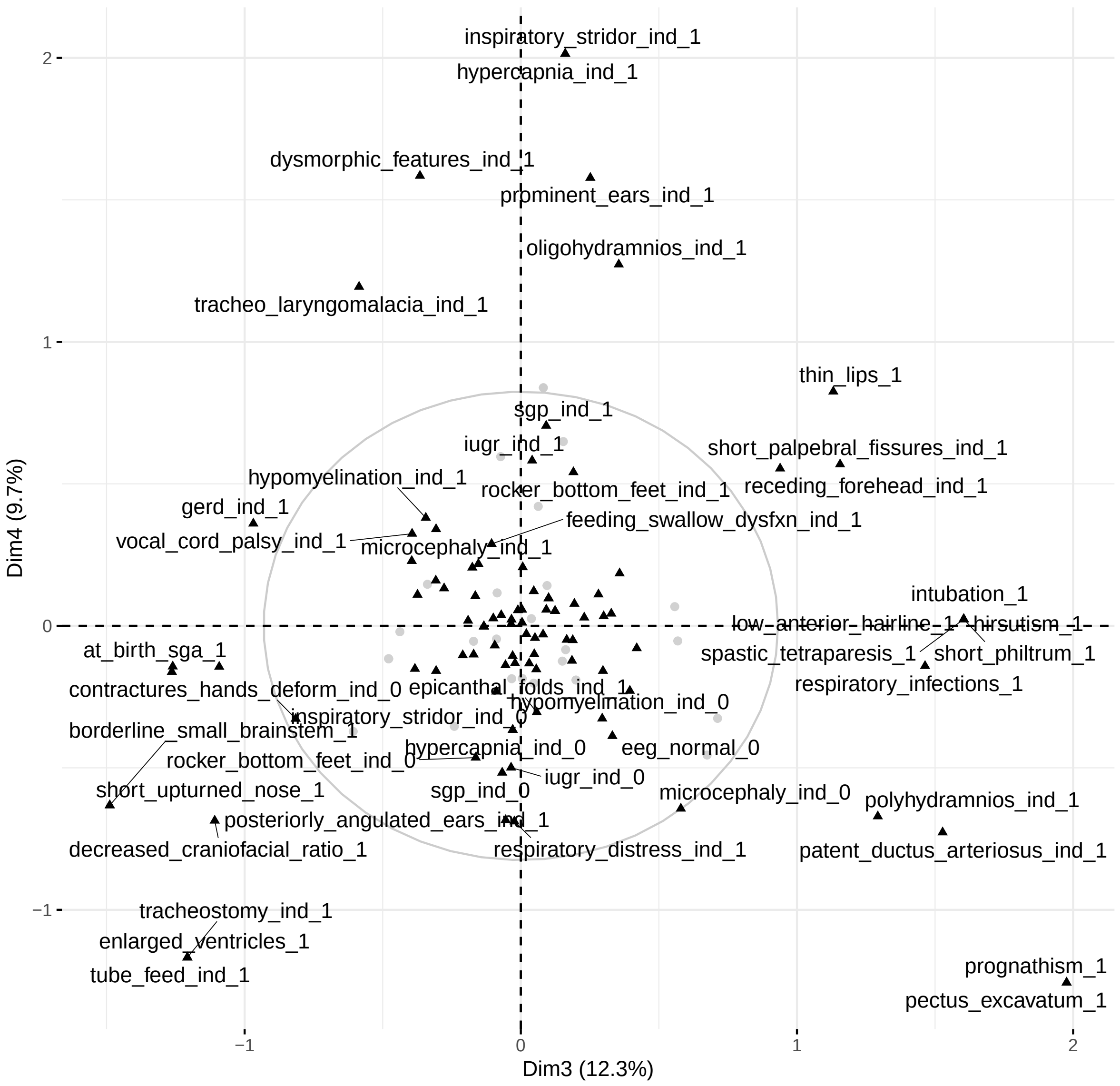


Figure 4: A bi-plot of principal axes 3 and 4, with grey cicles indicating the individual's projected position and black triangles indicaating the variables

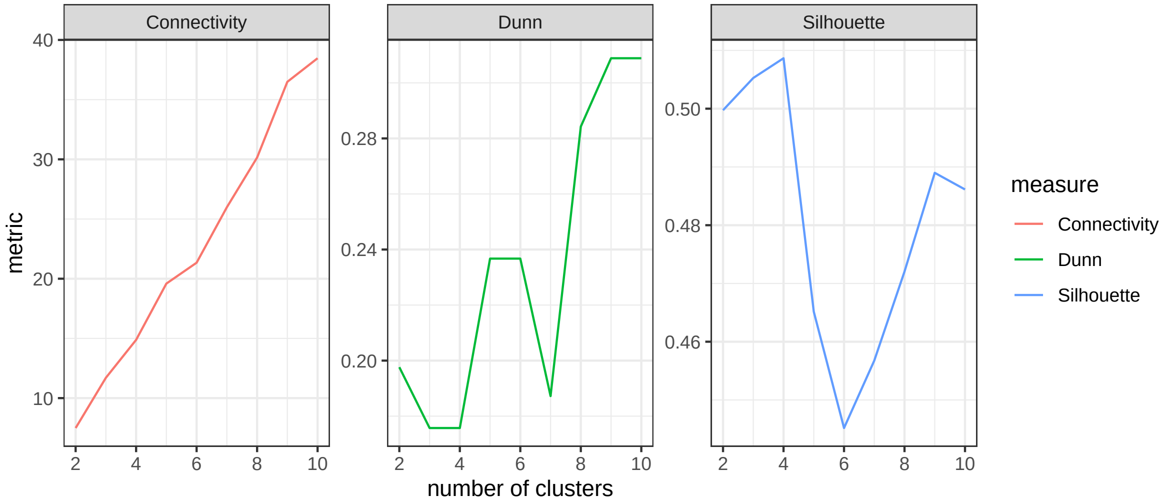


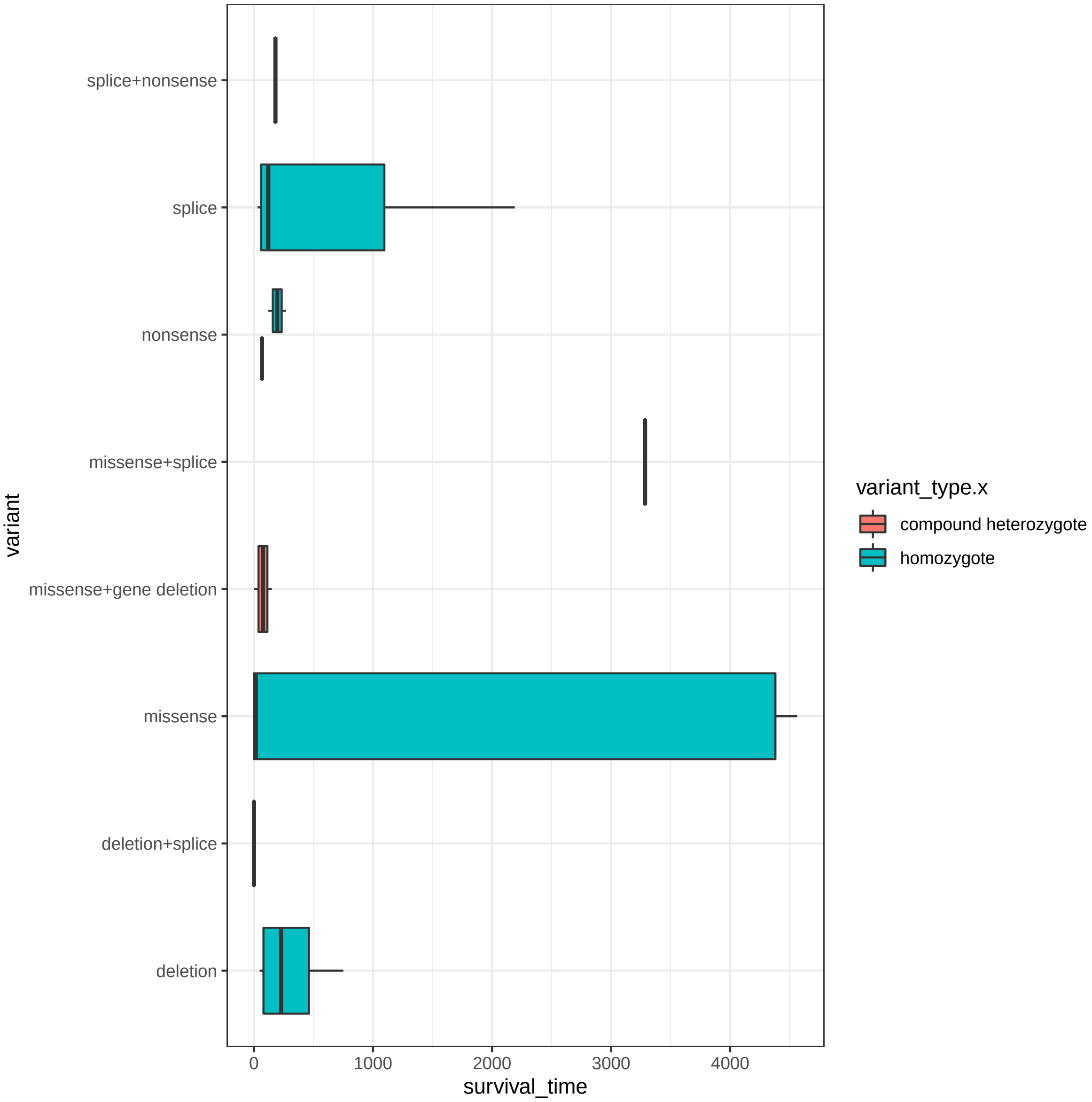
Figure 5: cluster validation analysis showing various metrics for cluster evaluation from 2 to 10 clusters. The most optimal number of clusters are selected by picking the lowest value from one or many of these compariable metrics.

Table 6: Survival probabilities derived from Kaplan-Meier estimation.

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
0.0	28	5	0	0.82	0.09	0.98	0.69
15.0	23	1	0	0.79	0.10	0.95	0.65
30.0	22	2	0	0.71	0.12	0.90	0.57
47.0	20	1	0	0.68	0.13	0.88	0.53
60.0	19	1	1	0.64	0.14	0.85	0.49
67.0	17	1	0	0.61	0.15	0.82	0.45
90.0	16	0	1	0.61	0.15	0.82	0.45
120.0	15	2	0	0.52	0.18	0.75	0.37
150.0	13	1	0	0.48	0.20	0.72	0.33
180.0	12	1	0	0.44	0.22	0.68	0.29
240.0	11	1	0	0.40	0.24	0.64	0.25
270.0	10	0	1	0.40	0.24	0.64	0.25
365.0	9	0	1	0.40	0.24	0.64	0.25
750.0	8	1	0	0.35	0.27	0.60	0.21
1095.0	7	1	0	0.30	0.31	0.56	0.16
1825.0	6	0	1	0.30	0.31	0.56	0.16
2190.0	5	1	0	0.24	0.39	0.52	0.11
3285.0	4	0	2	0.24	0.39	0.52	0.11
4380.0	2	1	0	0.12	0.81	0.59	0.02

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
4562.5	1	1	0	0.00	Inf	NA	NA

A comparitive boxplot of survival times by variant type shows significant outliers for missense+splice and missense variations. However, even within missense hymozygous variants the survival time varies greatly. Other missense variants, such as heterozygous missense with gene deletion do not show the same increased survival time as indeed noted by Magini et al. (2019).



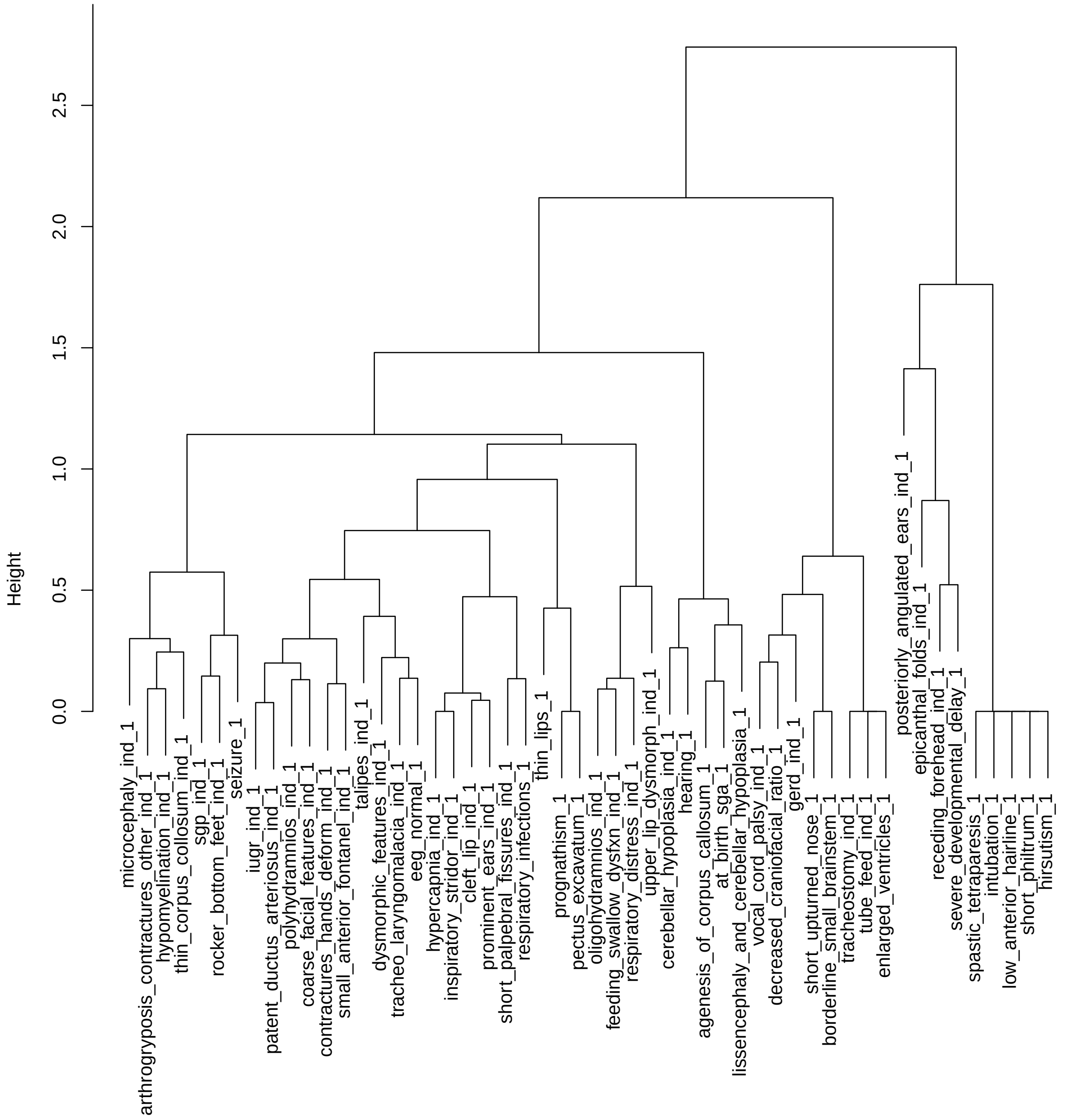


Figure 6: Dendrogram of heirarchical clustering method

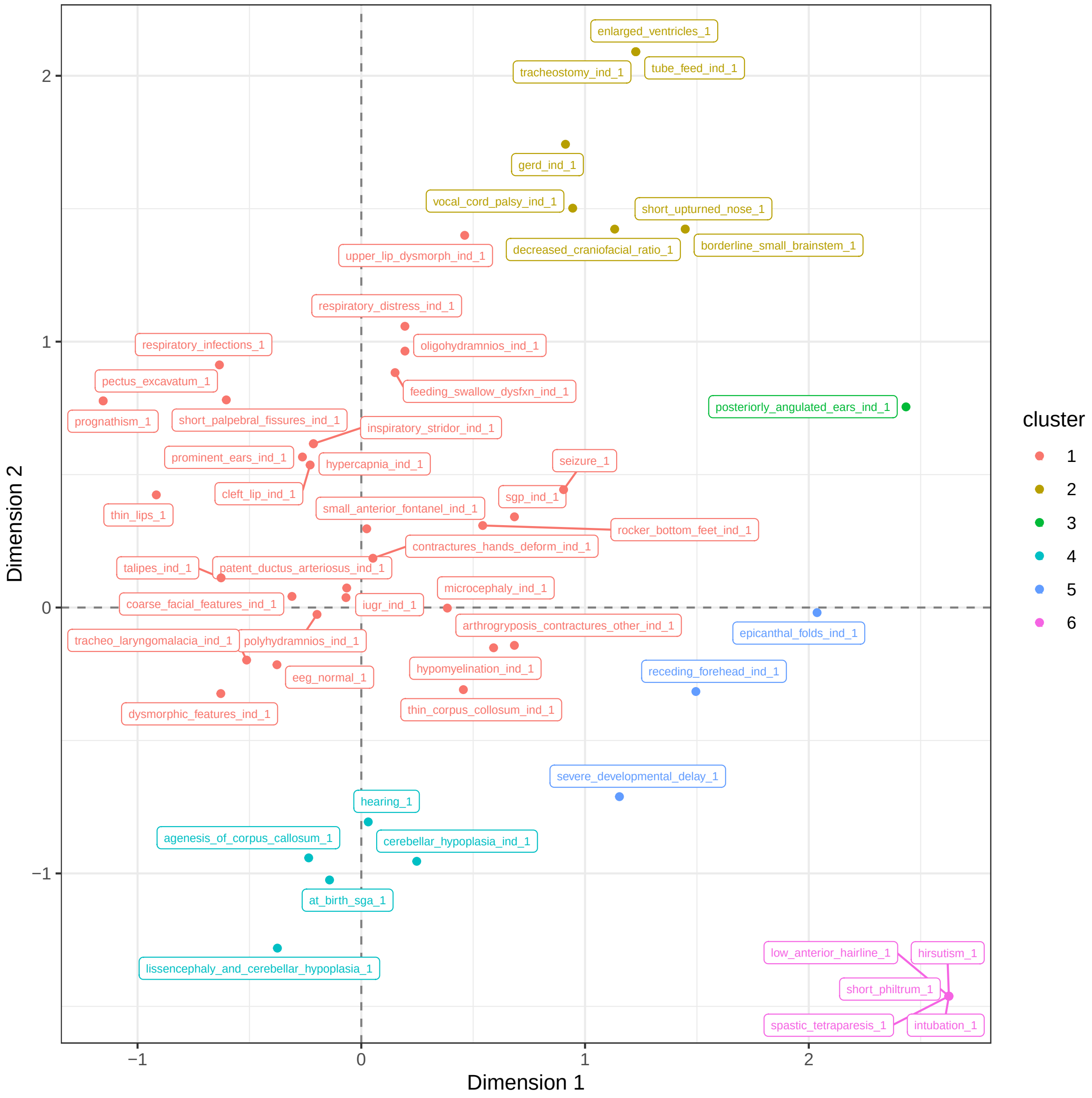


Figure 7: A projection of features on the first two principal components with cluster assignment indicated by colour.

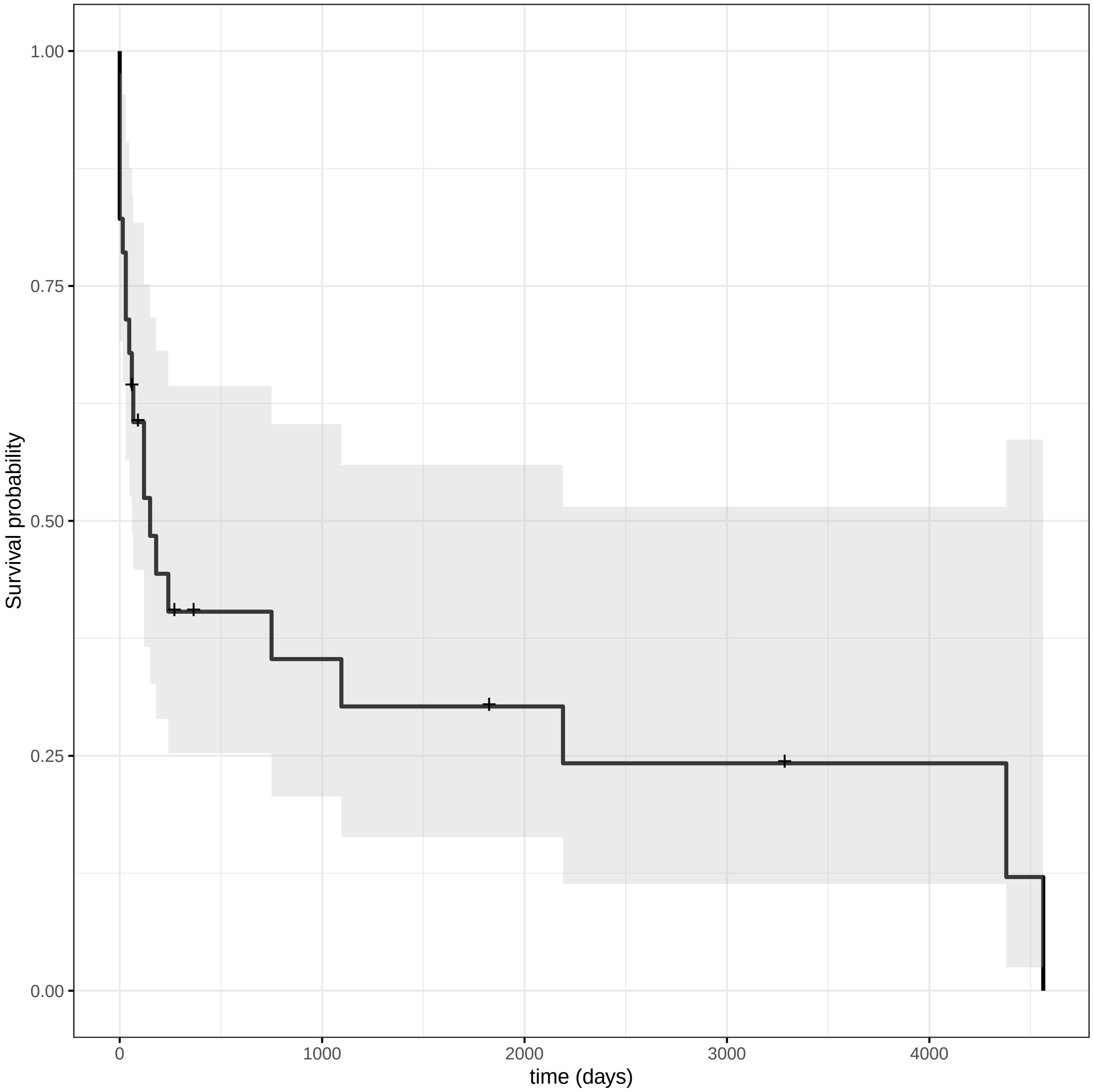


Figure 8: Kaplan Meier survival curve

The log rank test was carried out to compare the suvival curves of the various gene variants $\chi^2 = 14.8$ ($p = 0.04$) however due to the small sample size, several variants had expected counts less than 5 which would create a violation in assumptions for the test.

```
Call:
survdiff(formula = Surv(survival_time, deceased) ~ variant, data = survival_data)

                               N Observed Expected (O-E)^2/E (O-E)^2/V
variant=deletion               4        2    2.680   0.17263   0.21735
variant=deletion+splice        2        2    0.357   7.55714   9.55385
variant=missense               5        5    5.233   0.01035   0.02213
variant=missense+gene deletion 2        2    0.863   1.49694   1.73215
variant=missense+splice        2        0    2.654   2.65354   3.39942
variant=nonsense               3        2    1.941   0.00179   0.00218
variant=splice                 9        7    6.504   0.03781   0.06127
variant=splice+nonsense        1        1    0.768   0.07008   0.07839

 Chisq= 14.8  on 7 degrees of freedom, p= 0.04
```

Discussion

- the relationship between the results and the original hypothesis, i.e., whether they support the hypothesis, or cause it to be rejected or modified

The Multiple Correspondence Analysis confirmed the heterogeneous variation in the data with the first four princicple components being required to explain more than 50% of the variation in the dataset.

The examination of key contributing factors, as well as a visual inspection of the dat projected into this lower dimensional space, indicate the primary sources of variation in the clinical phenotype can be explained by few indivudals with very different clinical features. This has confounded the typical or classical presentation.

Beyond these outliers, a interesting disticntion is made between the hallmark features of microcephaly, hypomyelination, arthrogryposis, simplified gyal pattern and developmental delay with specific airway and swallowing features such as vocal cord paralysis, tracheostomy, GERD and tube feeding.

Further distictions are possible with more specific structural brainanomalies exhibiting as distinct sources of variation in the clininal phenotype.

Many of the outlying features identifes are specific facial or anatomical dysmophims. While a noted and core feature is around course facial features, those with a milder course had clinical reports in the available data that articulated these dysmophims in much more detail.

On a similar note, the reporting of ‘talipes’ as acondition seems distict as a source of variation when appearing as an isolated feature without the typical microcephasly and hypomyelination. However, there was significant heterogenity in how feet contractures were reported in clinical case reports. A distinction was made between reported cases with rocker bottom feed or congenital vertical talus and those with clubfeet/talipes. In practice its not clear how these diagnoses were arived at, and given the non-specific nature of this feature, care needs to be taken in distinguishing between similar manifesting features in the literaure where possible.

The cluster analysis does not aim to group individuals that are alike, but rather the clinical features themselves. This highlights common sources of variation on the clinical phenotype of the condition itself, rather than typical coincident features of any one child. While the number of clusters chosen all represent valid partition of the data, a larger number was selected to align the goals of the analysis which was to uncover distinct groups. A smaller number would have likely yeilded isolated outliers from the main body of the features. This still occurs, however more substructure is detected with the higher cluster number.

In terms of patient survival, the first few months show a sharp decline in survival probability confirming early demise in the infant period as a key factor. A small number of cases show much longer survival into the childhood and in some cases into the second decade of life. The current literature suggests the type of gene variant has an impact on surivival time, which seems to be supported in this work, however the limited sample size here prevents other meaningful analysis such as logrank tests or cox proprtional hazard modelling which are likely underpowered when comparing multiple groups.

Despite NEDMABA exhibting a diverse and hetergeneous clinical phenotype, it is possible to identify meaningful substructures in this data through the use of dimension reduction techniques (MCA) and cluster analysis. This has isolated the classic NEDMABA phenotype of microcephaly, hypomyelination, arthrogryposis, simplified gyal pattern and developmental delay. Furthermore it has isolated additional distinct groupings of features that are exhibited in patients such as vocal cord paralysis with tracheostomy, swallowing dysfunction, tube feeding and GERD. In additional to typical MRI findings of symplified gyral pattern and hypomyelination, there exists clusters of structural anomalies of the corpus calllosum and cereballa hypoplasia.

The survival of indivuduals with this condition is critical given early demise in infancy and childhood is a noted feature. This work identifies the survival probabilites and shows a one year survival rate of 0.403 (0.252, 0.644). A naive estimate, using just those number of deceased patients as proportion of the study population is 0.57. This highlights the importance of applying a survival analysis technique that incorporates censoring so as not to overestimate the survival probability.

Appendix

```
[1] 0.5714286
```

clust	var
1	microcephaly_ind_1
1	sgp_ind_1
1	iugr_ind_1
1	talipes_ind_1
1	rocker_bottom_feet_ind_1
1	contractures_hands_deform_ind_1
1	arthrogryposis_contractures_other_ind_1
1	hypercapnia_ind_1
1	cleft_lip_ind_1
1	inspiratory_stridor_ind_1
1	polyhydramnios_ind_1
1	coarse_facial_features_ind_1
1	oligohydramnios_ind_1
1	respiratory_distress_ind_1
1	prominent_ears_ind_1
1	upper_lip_dysmorph_ind_1
1	hypomyelination_ind_1
1	short_palpebral_fissures_ind_1
1	dysmorphic_features_ind_1
1	patent_ductus_arteriosus_ind_1
1	small_anterior_fontanel_ind_1
1	tracheo_laryngomalacia_ind_1
1	thin_corpus_collosum_ind_1
1	feeding_swallow_dysfxn_ind_1
1	seizure_1
1	thin_lips_1
1	prognathism_1
1	pectus_excavatum_1
1	respiratory_infections_1
1	eeg_normal_1
2	vocal_cord_palsy_ind_1
2	tracheostomy_ind_1
2	tube_feed_ind_1
2	gerd_ind_1
2	decreased_craniofacial_ratio_1
2	short_upturned_nose_1

clust	var
2	borderline__small__brainstem__1
2	enlarged__ventricles__1
3	posteriorly__angulated__ears__ind__1
4	cerebellar__hypoplasia__ind__1
4	hearing__1
4	agenesis__of__corpus__callosum__1
4	at__birth__sga__1
4	lissencephaly__and__cerebellar__hypoplasia__1
5	epicanthal__folds__ind__1
5	receding__forehead__ind__1
5	severe__developmental__delay__1
6	spastic__tetraparesis__1
6	intubation__1
6	low__anterior__hairline__1
6	short__philtrum__1
6	hirsutism__1

References

Andersen, Per Kragh, and Richard D Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics*, 1100–1120.

Bijarnia-Mahay, Sunita, Puneeth H Somashekar, Parneet Kaur, Samarth Kulshrestha, Vedam L Ramprasad, Sakthivel Murugan, Seema Sud, and Anju Shukla. 2022. "Growth and Neurodevelopmental Disorder with Arthrogryposis, Microcephaly and Structural Brain Anomalies Caused by Bi-Allelic Partial Deletion of Smpd4 Gene." *Journal of Human Genetics* 67 (3): 133–36.

Chen, Xintong, Xiaochen Sun, and Yujin Hoshida. 2014. "Survival Analysis Tools in Genomics Research." *Human Genomics* 8 (1): 1–5.

Costa, Patrício Soares, Nadine Correia Santos, Pedro Cunha, Jorge Cotter, and Nuno Sousa. 2013. "The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables in Healthy Ageing." *Journal of Aging Research* 2013.

Crowe, Francesca, Dawit T Zemedikun, Kelvin Okoth, Nicola Jaime Adderley, Gavin Rudge, Mark Sheldon, Krishnarajah Nirantharakumar, and Tom Marshall. 2020. "Comorbidity Phenotypes and Risk of Mortality in Patients with Ischaemic Heart Disease in the UK." *Heart* 106 (11): 810–16. https://doi.org/10.1136/heartjnl-2019-316091.

Díaz-Santiago, Elena, Fernando M Jabato, Elena Rojano, Pedro Seoane, Florencio Pazos, James R Perkins, and Juan AG Ranea. 2020. "Phenotype-Genotype Comorbidity Analysis of Patients with Rare Disorders Provides Insight into Their Pathological and Molecular Bases." *PLoS Genetics* 16 (10): e1009054.

Han, Lu, Susanne M Benseler, and Pascal N Tyrrell. 2018. "Cluster and Multiple Correspondence Analyses in Rheumatology: Paths to Uncovering Relationships in a Sea of Data." *Rheumatic Disease Clinics* 44 (2): 349–60.

Harrington, David P, and Thomas R Fleming. 1982. "A Class of Rank Test Procedures for Censored Survival Data." *Biometrika* 69 (3): 553–66.

Ji, Weigang, Xiangtian Kong, Honggang Yin, Jian Xu, and Xueqian Wang. 2022. "Case Report: Novel Biallelic Null Variants of Smpd4 Confirm Its Involvement in Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies." *Frontiers in Genetics* 13.

Kaplan, Edward L, and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81.

Le Roux, Brigitte, and Henry Rouanet. 2010. *Multiple Correspondence Analysis*. Vol. 163. Sage.

Lê, Sébastien, Julie Josse, and François Husson. 2008. "FactoMineR: A Package for Multivariate Analysis." *Journal of Statistical Software* 25 (1): 1–18. https://doi.org/10.18637/jss.v025.i01.

Magini, Pamela, Daphne J Smits, Laura Vandervore, Rachel Schot, Marta Columbaro, Esmee Kasteleijn, Mees van der Ent, et al. 2019. "Loss of Smpd4 Causes a Developmental Disorder Characterized by Microcephaly and Congenital Arthrogryposis." *The American Journal of Human Genetics* 105 (4): 689–705.

Marchiori, Dean. 2022. *Smpd4: Smpd4-Related Clinical Phenotype Data*. https:

//deanmarchiori.github.io/SMPD4.

Monies, Dorota, Mohammed Abouelhoda, Mirna Assoum, Nabil Moghrabi, Rafiullah Rafiullah, Naif Almontashiri, Mohammed Alowain, et al. 2019. "Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population." *The American Journal of Human Genetics* 104 (6): 1182–1201.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ravenscroft, Gina, Joshua S Clayton, Fathimath Faiz, Padma Sivadorai, Di Milnes, Rob Cincotta, Phillip Moon, et al. 2021. "Neurogenetic Fetal Akinesia and Arthrogryposis: Genetics, Expanding Genotype-Phenotypes and Functional Genomics." *Journal of Medical Genetics* 58 (9): 609–18.

Terry M. Therneau, and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.